D278–D283 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 1 December 2011*
doi:10.1093/nar/gkr1187

SitEx: a computer system for analysis of projections of protein functional sites on eukaryotic genes

Irina Medvedeva[1,*], Pavel Demenkov[1,2,3], Nikolay Kolchanov[1] and Vladimir Ivanisenko[1]

[1]Computer Proteomics Laboratory, Institute of Cytology and Genetics SB RAS, 10 Lavrentyeva Avenue, [2]Sobolev Institute of Mathematics SB RAS, 4 Acad. Koptyug Avenue and [3]Novosibirsk State University, 2 Pirogova Street, 630090 Novosibirsk, Russia

Received August 15, 2011; Revised October 2, 2011; Accepted November 15, 2011

ABSTRACT

Search of interrelationships between the structural–functional protein organization and exon structure of encoding gene provides insights into issues concerned with the function, origin and evolution of genes and proteins. The functions of proteins and their domains are defined mostly by functional sites. The relation of the exon–intron structure of the gene to the protein functional sites has been little studied. Development of resources containing data on projections of protein functional sites on eukaryotic genes is needed. We have developed SitEx, a database that contains information on functional site amino acid positions in the exon structure of encoding gene. SitEx is integrated with the BLAST and 3DExonScan programs. BLAST is used for searching sequence similarity between the query protein and polypeptides encoded by single exons stored in SitEx. The 3DExonScan program is used for searching for structural similarity of the given protein with these polypeptides using superimpositions. The developed computer system allows users to analyze the coding features of functional sites by taking into account the exon structure of the gene, to detect the exons involved in shuffling in protein evolution, also to design protein-engineering experiments. SitEx is accessible at http://www-bionet.sscc.ru/sitex/. Currently, it contains information about 9994 functional sites presented in 2021 proteins described in proteomes of 17 organisms.

INTRODUCTION

The functional sites of proteins are the major determinants of the biological function of proteins, being the most conserved portions of protein sequence. The features of structural organization of the protein functional sites at the level of gene exon structure may be consequential for evolutionary events such as formation of protein domain structure and those due to exon shuffling (1), the duplication of the domain and insertion of domain from one gene to another. A protein domain is often encoded by a single or several exons (2). The formation of mosaic proteins composed of several identical and non-identical domains has been demonstrated. Shuffling events of this kind might have significantly affected protein function and they have been considered as major mechanisms underlying forces operating during evolution and shaping the diversity of domain complexity (1).

Databases and program resources for projection of the domains and exon boundaries on amino acids sequences and spatial protein structures are currently available. Resources for projection of domains boundaries are both numerous and robust. These include the well-known databases such as Pfam (3), InterPro (4), TIGRFAMs (5), PROSITE (6), ProDom (7). Certain resources that describe the projections of exon and domain boundaries on primary and tertiary structure of proteins have been developed. Thus, web-resource XdomView (8) contains information on projection of exon and domain boundaries on 3D structure; SEDB (9) maps the exon boundaries and intron phases on multiple structural superimposition of 3D structures; ExDom (10) maps the location of exon boundaries and protein domains on sequence and spatial structure of proteins. However, information about functional sites is not contained in these resources.

A number of resources store data on the functional sites: Catalytic Site Atlas (11), scPDB (12), SitesBase (13), InterPro, PROSITE. They are constructed automatically or under expert control. Based on multiple alignment Gene3D could predict the active sites (14). Earlier we have developed the PDBSite database (15) that contains data on sites automatically extracted from PDB. It was integrated with the PDBSiteScan

*To whom correspondence should be addressed. Tel: +7(383) 363-49-24x1311; Fax: +7(383) 333-12-78; Email: brukaro@bionet.nsc.ru

program (16) for recognition of functional sites in protein spatial structure. However, these resources do not contain information on the boundaries of exons and domains.

The aim of this work was to develop a resource that describes projection of protein domains, functional sites and exon boundaries on protein and coding gene sequence.

In analysis of the structural–functional organization of proteins, search of sequence and spatial structure similarity among polypeptides encoded by individual exons appears promising. SitEx was integrated with the BLAST and 3DExonScan programs. BLAST is used to search similarity of a query protein to polypeptides encoded by individual exons stored in SitEx. 3DExonScan searches the similarity between 3D structures of query protein and polypeptides encoded by single exons. It could be helpful for search for the structural similarity between the polypeptides that do not necessarily possess the sequence similarity (17,18).

CONSTRUCTION AND CONTENT

The SitEx database provides data on mapped positions of exon and domain boundaries, also on functional sites in full-length protein and coding gene sequences. Here we examine only proteins with known 3D structure and corresponding genes. Figure 1 presents the SitEx structure and the main information tables in relational database constructed with MySQL 5.1. They are as follows. `Chain` gives information on the coding sequences and the proteins they encode, also on the identifiers from the Ensembl (19) database; `Site` contains information on protein functional sites, names, references to PDB (20), functional site description and some other information; `ExonPos` and `SitePos` provide information on positions of exon boundaries and functional site amino acids/codons in coding gene sequence and the corresponding protein. The other tables are designed for storing references to external databases, also for links between the tables.

Those protein structure, whose descriptions contained functional sites were selected from the PDB database. Their sequences and their encoding gene sequences, exon and domain borders (according to the Pfam classification) were extracted from the Ensembl database. To identify the amino acid residues in a functional site, the same approach was used as for PDBSite. Information on amino acid positions of functional site in the PDB sequence was retrieved from the SITE record. Information on the ligands, sequences, organisms, protein names and site identifiers was retrieved from PDB, too. The numberings of amino acid residues in PDB and Ensembl may differ. With this mind, we analyzed the sequence equivalence. Using the PDB identifier and ClustalW (21) alignment, the information on coding gene was retrieved from Ensembl (version of July 2009). Also, using ClustalW alignments, amino acid positions of functional site, exon, Pfam domains and SCOP structural region boundaries were identified in polypeptide chain. Information on folds was retrieved from the SCOP database (22). The programs retrieving the information from the external data and constructing SitEx database were developed in Perl. As a result of data integration, we retrieved 2021 non-redundant protein sequences that belong to 17 genomes of different organisms.

Search in SitEx using BLAST

Based on the data on exon boundaries in gene sequence stored in SitEx, two BLAST (23) databases were developed for sequences, exons and the polypeptides they encode. As a result, search of sequences involved in coding or in the formation of a protein functional site became available.

The 3DExonScan program

The 3DExonScan program aims for search of structural similarity with polypeptides that are encoded partly or entirely by a single exon. 3DExonScan uses the files in PDB format that are prepared for every such polypeptide, if the tertiary structure even for the part of it is known.

3DExonScan is based on the SSM and CE algorithms (24,25). The 3D protein structure is represented as elements of the secondary structure, an α-helix or a β-strand. This makes feasible quick comparisons of protein 3D structures with those in the SitEx database. This representation sets limitations on the length of polypeptide sequences.

3DExonScan ignores exons encoding polypeptides whose secondary structure consists of less than two elements. Many exons code for short protein regions that may possibly be only single α-helix, β-strand or loop. 3D structures of single α-helices or single β-strands from different proteins are very similar and structural superimposition will result, in most cases, in small RMSD. To avoid this, short exons that encode single helices or β-strands were excluded from analysis. Being aware of the limitations of the approach, we intend to use a strategy we developed earlier to speed up structural superimposition of functional sites in the PDBSiteScan.

Using the 3DExonScan program, the user obtains the structural homologues from database of 3D structures of polypeptides that are encoded by single exon. The results are presented as a table that contains reference to exon description in the SitEx, PDB identifier, protein description and the Z-score of structural superimposition. The threshold for the Z-score is set empirically at 3.2 (this corresponds to $P = 0.9993$). Superimpositions that have lower Z-score are eliminated by program. Having chosen the most suitable superimposition, the user gets a visualization of superimposed 3D structures of query protein and polypeptide with Jmol 3D structure viewer (26) (Figure 2). It is possible to download the structural superimposition in a PDB format.

SitEx content

A statistics page is provided. The database includes:

- 17 organisms (list of organisms is presented in Table 1): 75% of protein functional sites are

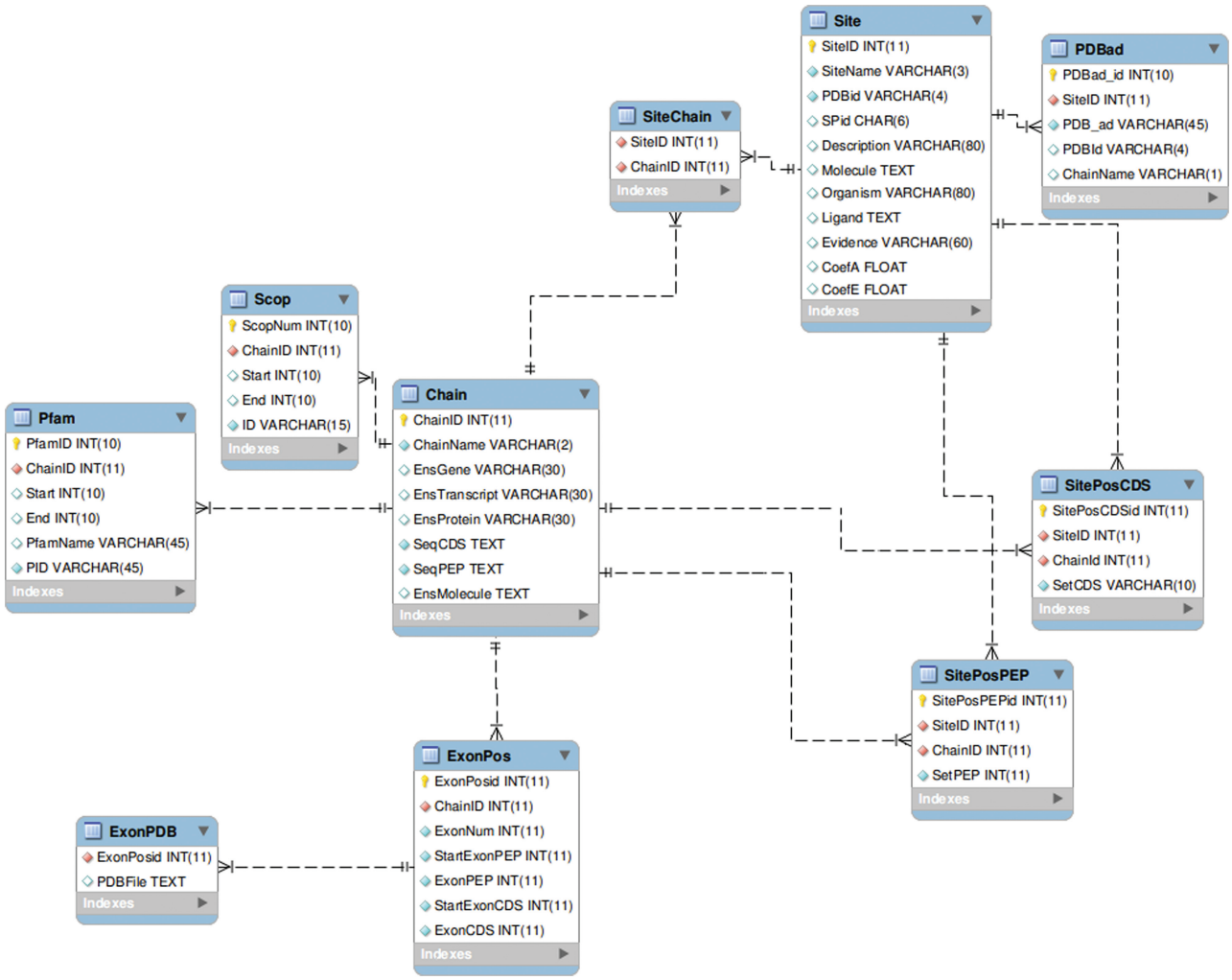

Figure 1. A schematic representation of the SitEx system. Headings correspond to table names. Table primary keys are in yellow. Foreign keys fields are in red. Type of field is next to its name. INT, numerical value; VARCHAR, character value; TEXT, text value.

represented by human proteins, 10% mouse, 5% rat, 5% bovine, 2% nematode, the remaining proteins occur singly;
- 715 ligands;
- 2021 non-redundant Ensembl protein sequences;
- 9994 (of 10 887) unique functional sites (sites that are different in amino acid composition and positions in similar PDB chains within one PDB entry);
- proteins were assigned to 13 groups according to their type; the enzyme group is widely represented (Table 2); and
- the functional sites are assigned to 14 groups according to the ligand type they bind (Table 3).

UTILITY

Interface

The web service SitEx consists of a database covering the protein functional sites, Pfam and SCOP domain projections, and provides search of the homologous exons and regions of the polypeptide chain encoded by a single exon (Exon BLAST Search page), and also search of the structurally similar polypeptides encoded by a single exon (3D Exon Search page).

One can access the information stored in SitEx through a user-friendly web-interface. The description page of the exon contains information on the exon identifiers, exon length, its boundaries, details concerning the coding protein according to the PDB and Ensembl data. An exon is in bold uppercase in both a coding sequence and an amino acid sequence. The functional sites are represented with uniquely colour background. Color symbols highlight sequence regions that belong to a single domain or structure (Figure 3A).

The description page of a functional site contains information on: (i) the protein that contains known functional site, the different chains on which a functional site is located according to the PDB data; (ii) the site amino acid positions in protein and in coding sequence; (iii) the discontinuity coefficients of a site on amino acid sequence and through exons; (iv) exon borders. A functional site is in bold uppercase in both a coding and an amino acid

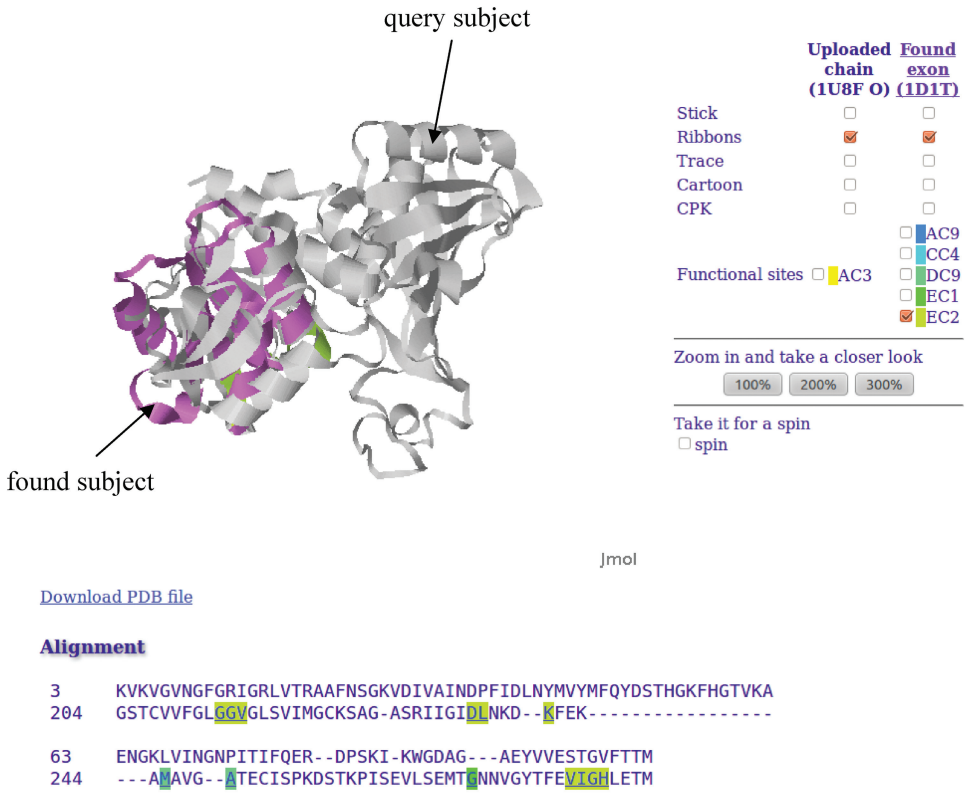

Figure 2. Superimposition of the 1U8F structure. Superimposition of the query 1U8F structure (D-glyceraldehyde-3-phosphate dehydrogenase, in grey) with the polypeptide encoded by the sixth exon of alcohol dehydrogenase that was found in SitEx database and incorporated into the 1D1T structure (alcohol dehydrogenase, in pink). The functional sites amino acids presented in found polypeptide were coloured on the sequence (the alignment of sequences presented for superimposed polypeptides is presented below) as well as on the 3D structure. The parameters of alignment are *Z*-score 3.9, RMSD 3.4. It seems to be a convergence due to the similar NAD-binding site because no sequence similarity was found even for exon sequences.

Table 1. List of organisms in SitEx

Organism
Rattus norvegicus
Mus musculus
Homo sapiens
Caenorhabditis elegans
Bos taurus
Gallus gallus
Oryctolagus cuniculus
Equus caballus
Xenopus tropicalis
Danio rerio
Anopheles gambiae
Ascaris suum
Sus scrofa
Echis multisquamatus
Cavia porcellus
Trimeresurus flavoviridis
Canis familiaris

Table 2. Protein classification in SitEx

Protein function	Number of entries
Muscle proteins	164
Blood proteins	21
Cell cycle proteins	245
Enzymes (with EC number, kinases, synthases)	2069
Immune system proteins	261
Membrane proteins	68
Receptors	213
Proteins participating in replication, transcription, translation	70
Heat shock proteins	20
Transport proteins	33
Tumour proteins	101
Zinc/RING fingers	131
Other proteins and precursors	618

sequences. The exons are represented with uniquely colour background. Color symbols highlight sequence regions that belong to a domain or a structure (Figure 3B).

There is the page for the current statistics for SitEx. Search queries are performed through the PDB, SCOP and Ensembl identifiers, names of the Pfam domains, organism, ligand, length and ordering of exon or any keywords (Figure 3C). Search of the exon length is based both on the number of nucleotides and that of amino acids encoded by the exon. A list of exons (and the polypeptides they code) together with the identification of encoded sites (Figure 3D) results from the query.

CONCLUSIONS

The SitEx system stores information on projection of exon and domain boundaries, positions of functional sites on protein sequence and the coding sequence of the gene.

This system provides opportunities for analysis of the relationships between the exon–intron structure of genes and functional sites of proteins. BLAST allows the search for the coding sequences and polypeptides encoding by them corresponding to single exon. 3DExonScan supplies the search through database of 3D structures of polypeptides encoded by single exon with structural superimposition. Using these programs, it is possible to investigate the sequence and structural conservation of the single exon coded polypeptides involved in the formation of a protein functional site.

SitEx is applicable in the study of the structural–functional organization of the gene and the features of coding and evolution of the functional sites by taking into account the exon structure of the gene; detection of exons involved in shuffling in an evolutionary perspective; it is helpful in rational design of novel proteins composed of fragments encoding individual exons from distinct genes.

In the nearest future we are going to expand the number of genomes represented in SitEx by fungi and plant Ensembl databases. During the next step, we intend to associate the functional site positions in gene encoding proteins with known 3D structure and exon boundaries of orthologous genes.

Table 3. Ligand classification in SitEx

Ligand class	Number of entries
Metal ions	2944
Acid anions	2434
Organic acids	551
Nucleotide phosphates	799
Sugars	308
Proteins	73
Amino acids and their compounds	164
Coenzymes	89
Alcohol and its compounds	665
Atoms and inorganic molecules	351
Amines and amides	1112
Porphyrins	59
Other ligands (alkaloids, ketones, pigments, etc.)	1354
Unknown ligands	396

FUNDING

Funding for open access charge: Ministry of Science & Education (grant numbers 14.740.11.0001, 07.514.11.4003, in part); Interdisciplinary Integrative Project 35 of SB RAS (94, 111, 119, in part); Russian Foundation for Basic Research (grant no. 11-04-92712, in part); FP7: EU-FP7 SYSPATHO No. 260429; Program of RAS (A.II.5, A.II.6, B.21, B.26, in part) and DAAD Leonard Euler Program Grant (in part).

Figure 3. Visualization of internal pages from SitEx. (**A**) exon description page. The start and end positions of the exon are shown. They correspond to the Ensembl numeration. AA, amino acid; bp, base pair; (**B**) a fragment of a page for description of functional site; (**C**) a query page; (**D**) result of query using the keyword 'ATP' in the ligand line. The functional site, information on the sequence, list of exons, which encode the sequence, their length measured in codons is displayed.

Conflict of interest statement. None declared.

REFERENCES

1. Todd,A.E., Orengo,C.A. and Thornton,J.M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, **307**, 1113–1143.
2. Kaessmann,H., Zollner,S., Nekrutenko,A. and Li,W. (2002) Signatures of domain shuffling in the human genome. *Genome Res.*, **12**, 1642–1650.
3. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunesekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
4. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
5. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
6. Sigrist,C.J.A., Cerutti,L., de Castro,E., Langendijk-Genevaux,P.S., Bulliard,V., Bairoch,A. and Hulo,N. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.*, **38**, 161–166.
7. Servant,F., Bru,C., Carrere,S., Courcelle,E., Gouzy,J., Peyruc,D. and Kahn,D. (2003) ProDom: automated clustering of homologous domains. *Brief. Bioinformatics*, **3**, 246–251.
8. Vivek,G., Tan,T.W. and Ranganathan,S. (2003) XdomView: protein domain and exon position visualization. *Bioinformatics*, **19**, 159–160.
9. Leslin,C.M., Abyzov,A. and Ilyin,V.A. (2004) Structural exon database, SEDB, mapping exon boundaries on multiple protein structures. *Bioinformatics*, **20**, 1801–1803.
10. Bhasi,A., Philip,P., Manikandan,V. and Senapathy,P. (2009) ExDom: an integrated database for comparative analysis of the exon-intron structures of protein domains in eukaryotes. *Nucleic Acids Res.*, **37**, D703–D711.
11. Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.
12. Kellenberger,E., Muller,P., Schalon,C., Bret,G., Foata,N. and Rognan,D. (2006) sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J. Chem. Inf. Model.*, **46**, 717–727.
13. Gold,N.D. and Jackson,R.M. (2006) SitesBase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Res*, **34**, D231–D234.
14. Lees,J., Yeats,C., Redfern,O., Clegg,A. and Orengo,C.A. (2010) Gene3D: merging structure and function for a Thousand genomes. *Nucleic Acids Res.*, **38**, D296–D300.
15. Ivanisenko,V.A., Pintus,S.S., Grigorovich,D.A. and Kolchanov,N.A. (2005) PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res.*, **33**, D183–D187.
16. Ivanisenko,V.A., Pintus,S.S., Grigorovich,D.A. and Kolchanov,N.A. (2004) PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucleic Acids Res.*, **32**, W549–W554.
17. Grishin,N.V. (2001) Fold change in evolution of protein structures. *J. Struct. Biol.*, **134**, 167–185.
18. Salemme,F.R., Miller,M.D. and Jordan,S.R. (1977) Structural convergence during protein evolution. *PNAS*, **74**, 2820–2824.
19. Flicek,P., Aken,B.L., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S. *et al.* (2010) Ensembl's 10th year. *Nucleic Acids Res.*, **38**, D557–D562.
20. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
21. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
22. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
23. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
24. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Cryst.*, **D6**, 2256–2268.
25. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
26. Herraez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.*, **34**, 255–261.